A Study Of Imputing Censored Observations For 2-Parameter Weibull Distribution Based On Random Censoring

Kuo-Ching Chiou

Department of Accounting and Statistics, The Overseas Chinese Institute of Technology

100 Chiao Kwang Rd., Taichung, 407, Taiwan, R.O.C.

Abstract : Censoring models are frequently used in reliability analysis to reduce experimental time. Three types of censoring models are type-I, type-II and random censoring. In this study, we focus on the right-random censoring model. In this model, if the failure time T exceeds its associated censoring time C, then the failure time becomes a censored observation (T^+) (see Miller (1981), Lawless (1982), Lee (1992), among others). Consider using the observed censoring time to impute the censored observation, which underestimates the true failure time. Tong and Chiou (2001) proposed two imputations of T_i^+ (parametric and non-parametric methods) and the author proposes another imputation of T_i^+ (non-parametric method). The three imputing methods aim to improve the underestimate of true failure time. In this paper, we consider the failure time to follow a 2-parameter Weibull distribution, because the Weibull distribution is widely used to model lifetimes. By a Monte Carlo simulation, these relative parameters, sample size n, censoring number r, shape parameter β, scalar parameter θ and number of replications N, are given, we apply the goodness of fit test (i.e. chi-square test) to compare the four methods of imputing censored observations. And then, we can obtain the best method of imputing censored observations such that the right-random censored data distribute to be closed to the original data (the Weibull distribution). For reliability analysis, the results can be provided for the parametric estimating, quantile estimating, and so on.

Keywords: *right-random censoring; failure time; censoring time; imputation value; Weibull distribution; goodness of fit test*

1. INTRODUCTION

Three censoring cases, type-I, type-II and random censoring, are frequently used in reliability analysis to save experimental time. Both type-I and type-II censoring cases are commonly used in engineering applications and the random censoring case is often employed in medical studies involving animals or clinical trials (see Miller (1981)).

In this study, we consider the right-random censoring case. The right-random censoring process is one in which an individual is assumed to have a failure time T and censoring time C, where T and C are independent continuous random variables. Assume that n individuals are considered and the *i*th individual has a failure time T_i and a censoring time C_i, for i = 1, 2, ..., n. Allow Y_1, Y_2, ..., Y_n to be the data from a right-random censoring case. Miller (1981), Lawless (1982), Lee (1992) and among others considered $Y_i = \min(T_i, C_i)$ for i = 1, 2, ..., n. Data from such a setup can be conveniently represented by the n pairs of random variables (Y_i, δ_i), where $\delta_i = 1$ if $T_i \le C_i$ and $\delta_i = 0$ if $T_i > C_i$ for i = 1, 2, ..., n. Restated, $Y_i = \delta_i T_i + (1 - \delta_i) C_i$, denotes whether the failure time T_i is censored or not; Y_i equals T_i if T_i is observed, and Y_i equals C_i if T_i is censored. Therefore, if the failure time T_i is a censored observation, denoted as T_i^+ (see Miller (1981)), then most of authors, Miller, Lawless, Lee, among others, considered the censoring time C_i to be an imputation of the censored observation T_i^+. Obviously, the T_i^+ is imputed by the censoring time C_i to underestimate the original failure time T_i. Chiou and Tong proposed two imputations of T_i^+ to improve the censoring time C_i. In this case, the pseudo random variables can be defined as $Y_i = T_i \cdot\delta_i + E(T_i \mid T_i > C_i) \cdot(1 - \delta_i)$ where $\delta_i = 1$ if $T_i \le C_i$ and $\delta_i = 0$ if $T_i > C_i$ for i = 1, 2, ..., n (see Buckly and James (1979)). For 2-parameter Weibull distribution, simulation results indicate that the two methods proposed by Chiou and Tong, herein were superior to the method, the censoring time imputation, if the shape parameter of Weibull distribution exceeds 1, except for the lower quantiles. The reason is that the data are distributed to be skew to the right for the shape parameter of Weibull distribution exceeds 1. The 2-parameter Weibull distribution is often used in life testing and reliability theory, because it models either increasing or decreasing failure rate in a sample manner. In this study, we consider that the failure time follow a 2-parameter Weibull distribution. And then compare the right-random censored data, by the four methods imputation censored observations, which distributes to original data (the Weibull distribution).

In this study, we consider four methods for imputing method of T_i^+ to collect the experimental data that the relative parameters, sample size n, censoring rate p (p = r / n, r is the number of the uncensored data), shape parameter β, scalar parameter θ and number of replications N, are given. By Monte Carlo simulation study, we use chi-square test (see Lee (1992)) to assess which data are approximated to the original data (the Weibull distribution).

2. DERIVING THE CONDITIONAL EXPECTATIONS

2.1. Deriving the approximation of conditional expectation for the empirical distribution

The conditional cumulative distribution function (c.d.f.) of a continuous random variable (r.v.) T given T > k is

$$\mathrm{P_r}\{T \le t \mid T > k\} = \frac{\mathrm{P_r}\{k < T \le t\}}{\mathrm{P_r}\{T > k\}} = \frac{\int_k^t f(x)dx}{1 - F(k)},$$

$t > k$.. (1)

The conditional probability density function (p.d.f.) of a continuous random variable (r.v.) T given T > k can be obtained by differentiating Eq.(1) with respect to t as follows:

$$f(t \mid T > k) = \frac{f(t)}{1 - F(k)}, \quad t > k. \quad (2)$$

The conditional expected value of a continuous r.v. T given T > k is

$$E(T \mid T > k) = \int_k^\infty t \cdot f(t \mid T > k)dt = \frac{\int_k^\infty t \cdot f(t)dt}{1 - F(k)} \quad (3)$$

For most reliability distributions, no simple closed form generally exists as (3). However, (3) can be approximated by a nonparametric empirical distribution as follows (see Chiou and Tong (2001)).

Let $t_{1:n}, t_{2:n}, \ldots, t_{n:n}$ be the ordered observations of $T_{1:n}, T_{2:n}, \ldots, T_{n:n}$, respectively, then

$$\int_{t_{i:n}^+}^\infty f(x)dx \approx \frac{n-i}{n} \quad (4)$$

and

$$\int_{t_{i:n}^+}^\infty x \cdot f(x)dx \approx \sum_{j=i+1}^{n} \frac{t_{j:n}}{n}. \quad (5)$$

By (4) and (5), we obtain

$$E(T \mid T > t_{i:n}) \approx \sum_{j=i+1}^{n} \frac{t_{j:n}}{n-i}. \quad (6)$$

2.2. Deriving the conditional expectation for 2-parameter Weibull distribution

If a r.v. T follows a 2-parameter Weibull distribution, then the p.d.f. of T is

$$f(t) = \frac{\beta}{\theta^\beta} t^{\beta-1} e^{-(t/\theta)^\beta}, \quad t > 0, \quad (7)$$

where the scale parameter θ and the shape parameter β are both positive.

The conditional p.d.f. and the expected value of T, given T > k, are

$$f(t \mid T > k) = \frac{\beta}{\theta^\beta} \cdot t^{\beta-1} \exp\left(-(t/\theta)^\beta + (k/\theta)^\beta\right), \quad t > k, \quad (8)$$

and

$E(T \mid T > k) =$

$\theta \cdot \exp((k/\theta)^{\beta}) \cdot \Gamma(1+1/\beta) \cdot [1 - I((k/\theta)^{\beta}, 1 + 1/\beta)]$, (9)

where $\Gamma(\gamma)$ and $I(\theta, \gamma)$ are the gamma function and the cumulative distribution function of Gamma distribution, respectively. The two functions can be defined as follows (see Lee (1992)):

$$\Gamma(\gamma)=\int_0^{\infty} u^{\gamma-1} \cdot e^{-u}\, du , \qquad (10)$$

and

$$I(\theta,\gamma) = \int_0^{\theta} \frac{1}{\Gamma(\gamma)} \cdot u^{\gamma-1} \cdot e^{-u}\, du. \qquad (11)$$

3. IMPUTATION OF CENSORED OBSERVATIONS BASED ON RIGHT-RANDOM CENSORING

In a right-random censoring, many authors (see Miller (1981), Lawless (1982), Lee (1992) and among others) consider the censoring time C_i is an imputation of the censored observation T_i^+. Instead, Chiou and Tong (2001) employed the pseudo random variables, $Y_i = T_i \cdot \delta_i + E(T_i \mid T_i > C_i) \cdot (1 - \delta_i)$ where $\delta_i = 1$ if $T_i \le C_i$ and $\delta_i = 0$ if $T_i > C_i$ for i = 1, 2, ..., n, which are utilized to construct the two imputations of T_i^+. The two imputations proposed herein can be obtained by substituting the values of $E(T_i \mid T_i > C_i)$ by the following two methods :

(1) Non-parametric method: By Eq.(6),

$$E(T_i \mid T_i > C_i) \approx \sum_{\substack{j=1 \\ j\neq i}}^{n} \{T_j \mid T_j \ge C_i \text{ and } T_j \text{ is}$$

an uncensored datum$\}/n_{ui}$, where n_{ui} denotes the number of $\{T_j \mid T_j \ge C_i$ and T_j represents an uncensored datum, j = 1, ... , n, $j \neq i\}$.

(2) Parametric method: Assuming that T_i follows a 2-parameter Weibull(θ, β) distribution, the conditional expected value $E(T_i \mid T_i > C_i)$ is equal to $\theta \cdot \exp[(C_i/\theta)^{\beta}] \cdot \Gamma(1 + 1/\beta) \cdot [1 - I((C_i/\theta)^{\beta}, 1 + 1/\beta)]$. The maximum likelihood estimates (MLE) of θ and β can be obtained by Keats *et al.* (1997) for the right-random censoring.

In this paper, the author proposes another imputation method that the censored observation T_i^+ is imputed by Median$\{T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n $\}$.

In right-random censoring, the experimental data, $Y_1, Y_2, \ldots, Y_n$, are collected. (i) $Y_i = \min(T_i, C_i)$ for i = 1, 2, ..., n, such method is denoted as "M1". (ii) $Y_i = T_i \cdot \delta_i + E(T_i \mid T_i > C_i) \cdot (1 - \delta_i)$ or $Y_i = T_i \cdot \delta_i +$ Median$\{T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n $\} \cdot (1 - \delta_i)$, where $\delta_i = 1$ if $T_i \le C_i$ and $\delta_i = 0$ if $T_i > C_i$ for i = 1, 2, ..., n. By non-parametric method, the two methods of imputations ($E(T_i \mid T_i > C_i) \approx \sum_{\substack{j=1 \\ j\neq i}}^{n} \{T_j \mid T_j \ge C_i$ and T_j is an uncensored datum$\} / n_{ui}$, and Median$\{T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n $\}$) are denoted as "M2" and "M4", respectively. By parametric method, the method is denoted as "M3" that the estimate

$$\hat{\theta}\exp[(C_i/\hat{\theta})^{\hat{\beta}}]\,\Gamma(1+1/\hat{\beta})\,[1-I((C_i/\hat{\theta})^{\hat{\beta}},1+1/\hat{\beta})]$$

of $E(T_i \mid T_i > C_i)$ is a imputation of the censored observation T_i^+. The estimates $\hat{\theta}$ and $\hat{\beta}$ are obtained by MLE (see Keats *et al.* (1997)) and the data, $Y_i = \min(T_i, C_i)$ for i = 1, 2, ..., n.

In right-random censoring, the experimental data, $y_1, y_2, \ldots, y_n$, are collected by four imputing methods, respectively. And then we use the chi-square test to assess which data distribute to be closed to the original distribution.

4. SIMULATION STUDY AND RESULTS

4.1. Simulation

Let the failure time T follow a Weibull(θ, β) distribution. A Monte Carlo simulation study was conducted to compare the performances of the four imputing methods ("M1", "M2", "M3" and "M4").

Note that the scale parameter θ is independent to suggest the best imputing method. So, the relative parameters are given as follows: sample size n = 30, 50, 100, censoring rate p = 0.1 (0.1) 0.5, shape parameter β = 0.1, 0.5, 1, 1.5, 2, 3, 4, 6 and scale parameter θ = 1.

For each combination of (n, p, β and θ), N = 1000 replications are generated by using IMSL STAT/LIBRARY (C Functions for Statistical Analysis). The simulation procedure is given below:

Step 1: Generate the data failure time T_i from a Weibull(θ, β) distribution, for i = 1, 2, ..., n.

Step 2: Determine how the censoring time C can be found under given p.

Let the failure time T follow a Weibull(θ, β), and a r.v. T* be $(T/\theta)^{\beta}$. By change of variable, we can derive the r.v. T* to follow a Weibull(1, 1). Let a r.v. C^* follow a Weibull(θ^*, 1). In right-random censoring, the failure time T and the censoring time C are independent random variables. So, θ^* can be found by the following equation.

$$p = P_r\{T > C\} = P_r\{(T/\theta)^{\beta} > (C/\theta)^{\beta}\} =$$

$$P_r\{T^* > C^*\} = \int_0^{\infty} (\int_z^{\infty} e^{-t} dt) \cdot \frac{1}{\theta^*} e^{\frac{-1}{\theta^*}} dz. \quad (12)$$

By the Eq.(12), θ^* = (1-p)/(pθ) = 1, θ = 1. So, θ^* is (1-p)/p. Next, we generate the data C^* from the Weibull(θ^*, 1) distribution, for i = 1, 2, ..., n. By transformation, the censoring time C_i is ${C^*_i}^{1/\beta}$, for i = 1, 2, ..., n.

Step 3: Accumulate the data Y_i for i = 1, 2, ..., n in a right-random censoring.

(i) $Y_i = \min(T_i, C_i)$ for i = 1, 2, ..., n.

(ii) $Y_i = T_i \cdot \delta_i + E(T_i \mid T_i > C_i) \cdot (1 - \delta_i)$ for i = 1, 2, ..., n. If $\delta_i = 0$ then $Y_i = E(T_i \mid T_i > C_i)$. The value of $E(T_i \mid T_i > C_i)$ is then replaced by method (1) (non-parametric method), method (2) (parametric method) in Section 3.1.

(iii) $Y_i = T_i \cdot \delta_i + \text{Median}\{T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n $\} \cdot (1 - \delta_i)$ for i = 1, 2, ..., n.

In this Step, the maximum datum of (Y_1, Y_2, ...,Y_n) must be constrained to be an uncensored datum. Otherwise, Eq. (6) can't be used and the data are discarded.

Step 4: Continue Step 3, the experimental data, y_1, y_2, …, y_n, are obtained by four imputing methods ("M1", "M2", "M3" and "M4"), respectively. By chi-square test, H_0: the experimental data, y_1, y_2, …, y_n, follow the Weibull(θ, β) distribution, that the parameters θ, β are given in step 1. By N=1000 replications, we can obtain the number that the conclusions do not reject H_0 for given the significant level $\alpha = 0.05$. And then, we suggest the best of imputing method that the number, the conclusions do not reject H_0, is the largest. Restated, the collected experimental data, the best imputing method, are the most closed to original data (the Weibull(θ, β)).

4.2. Results

As mentioned earlier in Section 3, the two imputations ($E(T_i \mid T_i > C_i)$ and Median$\{T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n $\}$) to impute censored observation are obviously larger than the imputation censoring time C_i. For censoring time, the defect of underestimation is obvious in shape parameter $\beta > 1$.

In Table 1, it proposes the simple suggestions of the best imputing method, and the suggestions as follow:

(1) For $\beta < 1$, we suggest to use imputing method "M1", except that the conditions, n =30 and p < 0.2, use the imputing method "M3".

(2) For $\beta \geq 1$, we suggest to use imputing method "M3" except p < 0.2.

(3) In n = 100 and p = 0.5, the number is too smaller that the conclusions do not reject H_0. In the conditions, we can not propose the suitable suggestions for imputing method, denoted as "*", because the censored

observations are too more to distribute more different from the original data (the Weibull distribution).

Table 1. The best imputing method for Weibull (1, β)

parameters	n = 30					n = 50					n = 100				
	P					p					p				
	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
β = 0.1	M3	M1	M1	M1	M1	M1	M1	M1	M1	M4	M1	M1	M1	M1	*
β = 0.5	M2	M3	M1	M1	M1	M3	M1	M1	M1	M1	M1	M1	M1	M1	*
β = 1	M2	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	*
β = 1.5	M2	M2	M3	M3	M2	M3	M3	M3	M3	M3	M2	M1	M1	M3	*
β = 2	M3	M2	M3	M3	M3	M3	M3	M3	M3	M3	M2	M3	M3	M3	*
β = 3	M2	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	M3	*
β = 4	M3	M2	M3	M3	M2	M3	M3	M3	M3	M3	M1	M3	M3	M3	*
β = 6	M3	M2	M3	M3	M2	M2	M3	M3	M3	M3	M1	M3	M3	M3	*

"": do not propose the best imputing method*

5. CONCLUSIONS

It is apparent that the censored observation is underestimated by censoring time in shape parameter β >1 for Weibull(θ, β). Chiou and Tong proposed that the censored observation T_i^+ was imputed by $E(T_i \mid T_i > C_i)$. And then, by the imputation $E(T_i \mid T_i > C_i)$, the estimates of moderate and high quantiles were superior to imputation censoring time C_i. In this study, the author intends whether the three imputing methods ("M2", "M3" and "M4") are superior to imputing method "M1" or not. By chi-quare test, the results show the three imputing methods ("M2", "M3" and "M4") are no inferior to imputing method "M1" (in Table 1).

For β < 1, the graph is a strictly decreasing for Weibull(θ, β). For censoring time, the defect of underestimation is not significant. So, the imputing method "M1" is superior than the other, except that the conditions, n =30 and p < 0.2. For β ≥1, the graph is a right-skewed for Weibull(θ, β). For censoring time, the defect of underestimation is significant. So, the imputing method "M3" is the superior than the other. In this study, the objective is that the results can provide the references for parametric estimating, quantile estimating, and so on.

For engineers, they get the experimental data in right-random censoring for imputations being censoring times. By goodness of fit test or hazard plot (see Lee (1992)), if the experimental data follow a 2-parameter Weibull distribution and the shape estimate β can be found, by hazard plot or MLE. Engineers can use this paper that it proposes the simple suggestions of the best imputing method to obtain the reliable data (see Table 1).

6. REFERENCES

Buckley, J. and James, I. (1979) Linear Regression with Censored Data. *Biometrika*, **66,** 429-436.

Chiou, K. C. and Tong, L. I. (2001) A Novel Means of Estimating Quantiles for 2-Parameter Weibull Distribution under the Right Random Censoring Model. *Advances and Applications in Statistics*, **1**, 1-26.

Keats, J. B., Lawrence, F. P. and Wang, F. K. (1997) Weibull Maximum Likelihood Parameter Estimates with Censored Data. *Journal of Quality Technology*, **29**, 105-110.

Lawless, J. F. (1982) *Statistical Models and Methods for Lifetime Data*. John Wiley, New York.

Lee, E. T. (1992) *Statistical Methods for Survival Data Analysis (2_{nd} ed)*. John Wiley and Sons, New York.

Miller, R. G. (1981) *Survival Analysis*. John Wiley and Sons, New York (1981).